Comparison of Robust Estimators for Detecting Outliers in Multivariate Datasets

Comparison of Robust Estimators for Detecting Outliers in Multivariate Datasets

Sharifah Sakinah Syed Abd Mutalib[1,2], Siti Zanariah Satari[1] and Wan Nur Syahidah Wan Yusoff[1]

[1]Centre for Mathematical Sciences, College of Computing and Applied Sciences, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Kuantan, Pahang, Malaysia

[2]Faculty of Computer, Media and Technology Management, University College TATI, Jalan Panchur, Telok Kalong, 24000 Kemaman, Terengganu, Malaysia

[1]Email: sharifahsakinah84@gmail.com

Abstract. Detecting outliers for multivariate data is difficult and does not work by visual inspection. Mahalanobis distance (MD) has been a classical method to detect outliers in multivariate data. However, classical mean and covariance matrix in MD suffer from masking and swamping effects. Masking effects happened when outliers are not identified and swamping effects happened when inliers are identified as outliers. Hence, robust estimators have been proposed to overcome these problems. In this study, the performance of a new robust estimator named Test on Covariance (TOC) is tested and compared with other robust estimators which are Fast Minimum Covariance Determinant (FMCD), Minimum Vector Variance (MVV), Covariance Matrix Equality (CME) and Index Set Equality (ISE). These five robust estimators' performance is being tested on five real multivariate datasets. Brain and weight, Hawkins-Bradu Kass, Stackloss, Bushfire and Milk datasets were used as these five real datasets are well-known in most outlier detection studies. Results show that TOC has proven to be able in detecting outliers, does not have a masking effect and has the same performance as other robust estimators in all datasets.

1. Introduction

The presence of outliers in multivariate data can affect the proper classical multivariate analysis, misleads the conclusions, make modeling difficult and disrupt measures of mean and covariance matrix. Outliers could be easily detected in univariate and bivariate data by using graphical presentation. However, the detection is difficult when the dimension increase [1,2].

One of the ways to detect multivariate outliers is to calculate a distance from each point to the center of the data. This method is known as the distance-based method which is based on the Mahalanobis distance (MD) [3]. An outlier would then be a point with a distance larger than some cut-off value. MD is one of the important tools to detect outliers in multivariate data [4]. MD is given by

$$d_i(\bar{\mathbf{x}},\mathbf{S}) = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})}, \quad i = 1,2,....,n \tag{1}$$

where $d_i(\bar{\mathbf{x}},\mathbf{S})$ is the MD for *i*-th observation, $\bar{\mathbf{x}}$ is the sample mean and $\mathbf{S}$ is covariance matrix [3].

The $\bar{\mathbf{x}}$ and $\mathbf{S}$ used in equation (1) are classical estimators for the sample mean and covariance matrix and not robust. A small portion of outliers will affect the estimation of $\bar{\mathbf{x}}$ and $\mathbf{S}$. MD depends on the sample mean and covariance matrix which are subject to masking and swamping effects [3–5]. Masking occurs when some of the outliers are left unidentified (false negative) and swamping occurs when non-outlying data are mistakenly identified as outliers (false positive) [6].

The outlier detection problems and the disadvantages of classical estimators in contaminated data have become a motivation to study robust methods for estimating mean and covariance matrix. A robust method is designed specifically to be resistant to outliers [3]. The robust method aims to lessen the effect of outliers and allow the majority of data to determine the result of the analysis [7]. A robust estimate of mean and covariance matrices are then replaced and used in MD and will yield robust MD or robust distance that is less sensitive to outliers [3,8].

Various robust estimators had been proposed and developed in the previous studies such as S, M, MM, Minimum Volume Ellipsoid (MVE), Minimum Covariance Determinant (MCD) and Fast-MCD (FMCD) estimators. Among these robust estimators, FMCD that had been developed by [9] is widely used because FMCD possesses the desirable properties of robust estimators which are affine equivariant, high breakdown point, bounded influence function and has lower computational complexity [2,10–12].

However, FMCD still has weaknesses which are the computational complexity as the dimension increase and singularity problems as FMCD is based on covariance determinant [2]. Therefore, in 2007, [2] proposed Minimum Vector Variance (MVV) to overcome the problem of FMCD. MVV can overcome the singularity problem as the computation of MVV is based on vector variance [2]. MVV was also found to have the same breakdown point as the MCD-based methods, lower computational time than FMCD, covariance matrix does not need to be positive definite and can be applied to high dimension datasets [2].

Although MVV can solve problems faced by FMCD, the computational time of MVV is still low when the number of variables increases [13]. Hence, [14] proposed Covariance Matrix Equality (CME) and Index Set Equality (ISE). The CME and ISE are the tests of equality between two covariance structures. These two robust estimators can find the robust mean and covariance matrices [15]. It is also found that ISE is simple to compute and has better performance than FMCD, MVV and CME [15].

However, ISE does not involve any arithmetical computation and it is still open to finding the test of equality between two covariance structures [13]. This motivates [16] to propose the test of equality between two covariance structures and named the new robust estimator with Test on Covariance (TOC). Details of TOC are discussed in [16].

A simulation study had been done in [16] to investigate the performance of TOC. Results from [16] show that TOC is applicable and a promising approach to detect outliers for multivariate data. Hence, in this study the performance of TOC will be investigating further by using real multivariate datasets. Multivariate datasets that will be used are Brain and Weight, Hawkins-Bradu Kass, Stackloss, Bushfire and Milk datasets. These five datasets had been used in most of the multivariate outlier detection literature and had become a benchmark to measure the performance of the proposed methods. The performance of TOC will be compared with other robust estimators (FMCD, MVV, CME and ISE) to detect outliers in these five datasets.

2. Robust Estimators

MVV, CME, ISE and TOC are modifications of the FMCD estimator. All these estimators differed at Step 6 in the FMCD algorithm. FMCD algorithm is given as follows [15].

Step 1: Select an arbitrarily subset H_{old} containing h different observations, where h is the smallest integer $\geq (n+p+1)/2$, where p is the number of variables and n is the sample size.

Step 2: Compute the mean vector $\bar{X}_{H_{old}}$ and covariance matrix $S_{H_{old}}$ of all observations belonging to H_{old}.

Step 3: Compute $d^2_{H_{old}}(i) = \left(X_i - \bar{X}_{H_{old}}\right)' S^{-1}_{H_{old}} \left(X_i - \bar{X}_{H_{old}}\right)$ for $i = 1,2,\ldots,n$.

Step 4: Sort $d^2_{H_{old}}(i)$ for $i=1,2,\ldots,n$ in increasing order $d^2_{H_{old}}(\pi(1)) \le d^2_{H_{old}}(\pi(2)) \le \ldots \le d^2_{H_{old}}(\pi(n))$ where π is a permutation on $\{1,2,\ldots,n\}$.

Step 5: Define $H_{new} = \{X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(h)}\}$ and then calculate $\bar{X}_{H_{new}}$, $S_{H_{new}}$ and $d^2_{H_{new}}(i)$ for $i=1,2,\ldots,n$.

Step 6_{FMCD}: If $\det(S_{H_{new}}) = 0$, repeat Step 1 – Step 5. Otherwise, if $\det(S_{H_{new}}) < \det(S_{H_{old}})$, let $H_{old} := H_{new}$, $\bar{X}_{H_{old}} := \bar{X}_{H_{new}}$ and $S_{H_{old}} := S_{H_{new}}$. Then go to Step 3. Otherwise, the process is stop and $\det(S_{H_{new}}) = \det(S_{H_{old}})$ is obtain.

The procedures of MVV, CME and ISE can be done by replacing Step 6_{FMCD} with the following step as given in [13,15].

Step 6_{MVV}: If $Tr(S^2_{H_{new}}) = 0$, repeat Step 1 – Step 5. Otherwise, if $Tr(S^2_{H_{new}}) \neq Tr(S^2_{H_{old}})$, let $H_{old} := H_{new}$, $\bar{X}_{H_{old}} := \bar{X}_{H_{new}}$ and $S_{H_{old}} := S_{H_{new}}$. Then go to Step 3. Otherwise, the process is stopped and $Tr(S^2_{H_{new}}) = Tr(S^2_{H_{old}})$ is obtained.

Step 6_{CME}: If $\sqrt{Tr(S_{H_{new}} - S_{H_{old}})^2} \neq 0$, calculates $\bar{X}_{H_{new}}$ and let $H_{old} := H_{new}$, $\bar{X}_{H_{old}} := \bar{X}_{H_{new}}$ and $S_{H_{old}} := S_{H_{new}}$. Then go to Step 3. Otherwise, the process is stop.

Step 6_{ISE}: If $I_{new} \neq I_{old}$, let $H_{old} := H_{new}$, calculate $\bar{X}_{H_{new}}$ and let $H_{old} := H_{new}$, $\bar{X}_{H_{old}} := \bar{X}_{H_{new}}$ and $S_{H_{old}} := S_{H_{new}}$. Then go to Step 3. Otherwise, the process is stop.

A new robust estimator named TOC had been proposed by [16]. The idea of TOC is coming from CME and ISE which test the equality of covariance structure for old subset and new subset in the algorithm. The equality of two covariance structures is tested by using equation (2) with the hypothesis $H_0 : \Sigma_{old} = \Sigma_{new}$ versus $H_1 : \Sigma_{old} \neq \Sigma_{new}$.

$$u = \upsilon\left[\sum_{i=1}^{p}(\lambda_i - \ln\lambda_i) - p\right] \tag{2}$$

where $\upsilon = n-1$, p is the number of variables and $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the eigenvalues of $\Sigma_{new}\Sigma^{-1}_{old}$. H_0 is rejected if $u > \chi^2\left[\alpha, \frac{1}{2}p(p+1)\right]$ as given in [17]. Step 6 for TOC is given below.

Step 6_{TOC}: If H_0 is rejected, calculate $\bar{X}_{H_{new}}$ and let $H_{old} := H_{new}$, $\bar{X}_{H_{old}} := \bar{X}_{H_{new}}$ and $S_{H_{old}} := S_{H_{new}}$. Then go to Step 3. Otherwise, the process is stop.

3. Illustrative Examples and Performance Measures

In this study, five real multivariate datasets will be used as illustrative examples to identify outliers in multivariate data. The datasets are Brain and Weight, Hawkins-Bradu Kass, Stackloss, Bushfire and Milk datasets. These datasets have become a standard for most of outlier detection studies in multivariate data such as in [1], [18 – 22]. Table 1 shows a summary of the datasets.

Table 1. Summary of the datasets.

No.	Data set	*n*	*p*	Number of Outliers	Percentage of outliers	Remarks
1	Brain and weight data (BW) dataset	28	2	3	11%	Observations 6th, 16th and 25th are outliers [3,18,23,24].
2	Hawkins-BraduKass (HBK) dataset	75	3	14	19%	Observations 1-14 are outliers [1,18–20,25,26].
3	Stackloss dataset	21	3	4	19%	Observations 1st – 3rd and 21st are outliers [19].
4	Bushfire dataset	38	5	13	34%	Observations 7th – 11th and 31st – 38th are outliers [20,21,27,28].
5	Milk dataset	86	8	17	20%	Observations 1st - 3rd, 12th, 13th - 17th, 27th, 41st, 44th, 47th, 70th, 74th, 75th and 77th are outliers [20,21,29].

Robust mean and covariance matrices from FMCD, MVV, CME, ISE and TOC will be obtained and were used to identify outliers in the datasets. The steps to identify outliers are given as follows,

Step 1: Compute the distance $d_i(\bar{\mathbf{x}},\mathbf{S})=\sqrt{(\mathbf{x}_i-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})}$ for $i=1,2,\ldots,n$.

Step 2: Use the cut-off value $\sqrt{\chi^2_{p,0.975}}$ to detect outliers. If $d(\mathbf{x}_i,\mathbf{S})>\sqrt{\chi^2_{p,0.975}}$, $\mathbf{x}_i$ is an outlier.

The performance of each robust estimator for each dataset will be measured by three measurements.

i. Number of outliers successfully detected.
The number of outliers detected by each robust estimator will be counted. Observations that are outliers from each dataset have already been identified in the previous study (refer to Table 1). Each robust estimator will be investigated either it can detect the outliers or not.

ii. Number of outliers falsely detected as inliers (masking effect).
Any outliers that are not identified as outliers will be counted as having a masking effect. Each robust estimator will be investigated either it misclassifies outliers as inliers.

iii. Number of inliers falsely detected as outliers (swamping effect).
Any inliers that are not identified as inliers will be counted as having a swamping effect. Each robust estimator will be investigated either it misclassifies inliers as outliers.

4. Results and Discussion

In this section, we compare and discuss the performance of FMCD, MVV, CME, ISE and TOC on multivariate outlier detection by using five real multivariate datasets.

4.1 Brain and Weight data

Brain and weight (BW) dataset contain two variables which are body weight and brain weight for 28 species of animals. According to [18] and [19], this dataset is part of a larger dataset in [30]. [1] used Minimum Volume Ellipsoid (MVE) in their study and found that observations 6th,14th, 16th, 17th and 25th are outliers. Observations 6th, 16th and 25th are dinosaurs with a small brain and heavy body, while observations 14th and 17th are human and rhesus monkey with high brain weight [1]. However, the method used by [1] tends to detect too many outliers [31]. According to [18,19,23], it is believed that this dataset only has three outliers which are observations 6th, 25th and 16th.

Table 2 shows the results for each robust estimator. At the cut-off value of $\sqrt{\chi^2_{2,0.975}} = 2.716$, all five robust estimators successfully detect outliers in the datasets and do not misclassify outliers as inliers. This shows all robust estimators do not have a masking effect for this dataset. However, all robust estimators misclassified observations 2nd, 7th, 14th, 15th and 24th as an outlier and hence have a swamping effect.

Table 2. Results for Brain and Weight dataset.

Performance Measures	FMCD	MVV	CME	ISE	TOC
Number of outliers successfully detected	3 (100%)	3 (100%)	3 (100%)	3 (100%)	3 (100%)
Number of outliers falsely detected as inliers (masking effect)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Number of inliers falsely detected as outliers (swamping effect)	5 (20%)	5 (20%)	5 (20%)	5 (20%)	5 (20%)

4.2 Hawkins-Bradu Kass (HBK) data

Hawkins-Bradu-Kass (HBK) dataset is an artificial dataset generated by [32]. This dataset was generated to show some of the merits of robust methods and the effectiveness of the robust methods to identify outliers [18]. This dataset has 75 observations and four variables (one response and three explanatory variables). For this study, only three explanatory variables will be used. Observations 1 – 14 are known to be outliers for this dataset [1,18–20].

By using the cut-off value of $\sqrt{\chi^2_{3,0.975}} = 3.058$, all robust estimators are successfully detect the outliers and do not misclassify outliers as inliers. All robust estimators do not have a masking effect for this dataset. For the swamping effect, only MVV misclassified observation 53rd as an outlier.

Table 3. Results for Hawkins-Bradu Kass dataset.

Performance Measures	FMCD	MVV	CME	ISE	TOC
Number of outliers successfully detected	14 (100%)	14 (100%)	14 (100%)	14 (100%)	14 (100%)
Number of outliers falsely detected as inliers (masking effect)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Number of inliers falsely detected as outliers (swamping effect)	0 (0%)	1 (1.6%)	0 (0%)	0 (0%)	0 (0%)

4.3 Stackloss data

Stackloss data is a dataset obtained from an experiment for the oxidation of ammonia into nitric acid, measured on 21 consecutive days [22]. The dataset has three explanatory variables (rate of incoming ammonia, cooling water temperature and acid concentration) and one response variable (stackloss) [19,22]. In this study, only three explanatory variables are used. Observations 1st, 2nd, 3rd and 21st are outliers [18,19].

From Table 4, all robust estimators successfully detect the outliers and do not misclassify outliers as inliers at the cut-off value of $\sqrt{\chi^2_{3,0.975}} = 3.058$. All robust estimators do not have a masking effect for this dataset as well. However, all robust estimators misclassified observation 17th as an outlier.

Table 4. Results for Stackloss dataset.

Performance Measures	FMCD	MVV	CME	ISE	TOC
Number of outliers successfully detected	4 (100%)	4 (100%)	4 (100%)	4 (100%)	4 (100%)
Number of outliers falsely detected as inliers (masking effect)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Number of inliers falsely detected as outliers (swamping effect)	1 (5.9%)	1 (5.9%)	1 (5.9%)	1 (5.9%)	1 (5.9%)

4.4 Bushfire data

Bushfire data is a dataset used to locate bushfire scars and was taken from [33]. The dataset contains satellite measurements on five frequency bands, corresponding to each of 38 pixels with 13 outliers which makes the percentage of outliers is 34%. According to [20,21], observations 7^{th} – 11^{th} and 31^{st} – 38^{th} are classified as outliers.

Only FMCD, ISE and TOC have successfully detected all outliers in Bushfire dataset. MVV only can detect 84.6% of outliers, while CME can only detect 53.8% of outliers. Therefore, MVV and CME misclassify outliers as inliers (masking effect) with the rate of 15.4% and 46.2% respectively. Table 5 shows that all robust estimators misclassify inliers as outliers (swamping effect). CME has the highest percentage which is 40%, while FMCD only misclassifies 8% of inliers as outliers. Meanwhile, TOC and ISE has a similar performance.

Table 5. Results for Bushfire dataset.

Performance Measures	FMCD	MVV	CME	ISE	TOC
Number of outliers successfully detected	13 (100%)	11 (84.6%)	7 (53.8%)	13 (100%)	13 (100%)
Number of outliers falsely detected as inliers (masking effect)	0 (0%)	2 (15.4%)	6 (46.2%)	0 (0%)	0 (0%)
Number of inliers falsely detected as outliers (swamping effect)	2 (8%)	5 (20%)	10 (40%)	3 (12%)	3 (12%)

4.5 Milk data

Milk dataset provided by [34] is a composition of 86 containers of milk with 8 variables. The 8 variables are density, fat content, protein content, casein content, cheese dry substance measured in factory, cheese dry substance measured in laboratory, milk dry substance and cheese produced. There are 17 outliers in this dataset which makes the percentage of outliers is 20%. Observations 1^{st}-3^{rd}, 12^{th}-17^{th}, 27^{th}, 41^{st}, 44^{th}, 47^{th}, 70^{th}, 74^{th}, 75^{th} and 77^{th} are classified as outliers by [20] , [21] and [29].

As can be seen from Table 6, all robust estimators successfully detect 17 outliers and do not misclassify outliers as inliers (masking effect). However, all robust estimators have a swamping effect. FMCD, MVV and ISE show the lowest swamping effect with 5.8%, while CME and TOC have 7.2%.

Table 6. Results for Milk dataset.

Performance Measures	FMCD	MVV	CME	ISE	TOC
Number of outliers successfully detected	17 (100%)	17 (100%)	17 (100%)	17 (100%)	17 (100%)
Number of outliers falsely detected as inliers (masking effect)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Number of inliers falsely detected as outliers (swamping effect)	4 (5.8%)	4 (5.8%)	5 (7.2%)	4 (5.8%)	5 (7.2%)

Table 7 shows a summary of the best robust estimators for each dataset. It can be seen that all robust estimators successfully detected outliers for four datasets (BW, HBK, Stackloss and Milk). For Bushfire dataset, only FMCD, ISE and TOC successfully detected all outliers. The same results can be seen for the masking effect. All robust estimators do not have a masking effect for four datasets except for the Bushfire dataset. Only FMCD, ISE and TOC do not misclassify outliers as inliers (masking effect) for Bushfire dataset. For the swamping effect, all robust estimators misclassify inliers as outliers for four datasets (BW, Stackloss, Bushfire and Milk) except HBK dataset. For HBK dataset, only MVV has a swamping effect.

Table 7. Summary of the best robust estimators for five datasets.

Datasets	All outliers successfully detected	No outliers falsely detected as inliers (No masking effect)	No inliers falsely detected as outliers (No swamping effect)
Brain and weight (BW)	All	All	None
Hawkins-BraduKass (HBK)	All	All	FMCD, CME, ISE, TOC
Stackloss	All	All	None
Bushfire	FMCD, ISE, TOC	FMCD, ISE, TOC	None
Milk	All	All	None

5. Conclusions

In this study, the performance of a new robust estimator by [16] named Test on Covariance (TOC) to detect outliers in real multivariate datasets is being tested and compared with other robust estimators. The performance of TOC is compared with Fast Minimum Covariance Determinant (FMCD), Minimum Vector Variance (MVV), Covariance Matrix Equality (CME) and Index Set Equality (ISE). These five robust estimators' performance is being tested on five real multivariate datasets which are Brain and weight (BW), Hawkins-Bradu Kass (HBK), Stackloss, Bushfire and Milk datasets. The performance of each robust estimator measured by the number of outliers successfully detected, number of outliers falsely detected as inliers (masking effect) and number of inliers falsely detected as outliers (swamping effect).

Ideally, the best robust estimator would be a robust estimator that can detect all outliers, has the lowest masking and swamping effect. It is found in this study that all robust estimators successfully detected outliers in BW, HBK, Stackloss and Milk datasets. The same result was also obtained for the masking effect. However, only FMCD, ISE and TOC successfully detect outliers and do not have a masking effect for Bushfire dataset. For the swamping effect, all robust estimators misclassify inliers as outliers for BW, Stackloss, Bushfire and Milk dataset. Meanwhile FMCD, CME, ISE and TOC do not have the swamping effect of HBK dataset. This means that FMCD, CME, ISE and TOC did not misclassify inliers as outliers for HBK dataset. From these results, TOC has proven to be able to detect outliers, does not have a masking effect and has the same performance as other robust estimators in five real multivariate datasets. This shows that TOC is applicable and a promising approach for outlier detection in multivariate data. Hence, TOC can be used when outliers are existed in multivariate datasets.

Acknowledgment

The authors would like to thank University College TATI (UC TATI) for the financial support under the grant UC TATI Short Term Research Grant (STG) (9001-1905), Internal Grant (RDU 190363 and RDU 1901168) Universiti Malaysia Pahang (UMP) and Ministry of Higher Education Malaysia [FRGS/1/2019/STG06/UMP/02/6].

References

[1] Rousseeuw PJ and Van Zomeren BC 1990 Unmasking multivariate outliers and leverage points *J. Am. Stat. Assoc*. **85** 633 - 9.
[2] Herwindiati DE, Djauhari MA and Mashuri M 2007 Robust multivariate outlier labeling *Commun. Stat. Simul. Comput.* **36** 1287–94.
[3] Hadi AS, Rahmatullah Imon AHM and Werner M 2009 Detection of outliers. *Wiley Interdiscip. Rev. Comput. Stat.* **1** 57–70.
[4] Werner M 2003 *Identification of multivariate outliers in large data sets*.
[5] Rousseeuw PJ and Hubert M 2011 Robust statistics for outlier detection. *Wiley Interdiscip Rev Data Min. Knowl. Discov*. **1** 73–9.
[6] Filzmoser P and Todorov V 2013 Robust tools for the imperfect world. *Inf Sci (Ny)*. **245** 4–20.
[7] Møller SF, Von Frese J and Bro R 2005 Robust methods for multivariate data analysis. *J.Chemom.***19** 549–63.
[8] Su X and Tsai C-L 2011 Outlier detection *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **1** 261–8.
[9] Rousseeuw PJ and Van Driessen K 1999 A fast algorithm for the minimum covariance determinant estimator *Technometrics* **41** 212–23.
[10] Djauhari MA 2008 A robust estimation of location and scatter *Malaysian J. Math. Sci.* **2** 1–24.
[11] Djauhari MA 2008 *Highly robust estimation of location and scatter when data sets are of high dimension: An open problem* (The 3rd International Conference on Mathematics and Statistics (ICoMS-3)) p 1–8.
[12] Hubert M and Debruyne M 2010 Minimum covariance determinant *Wiley Interdiscip. Rev. Comput. Stat.* **2** 36–43.
[13] Salleh RM and Djauhari MA 2011 *Robust hotelling's T^2 control charting in spike production process* (Int Semin Appl Sci Math 2011 (ISASM 2011)) p 1–8.
[14] Rohayu S 2013 *A robust estimation method of location and scale with application in monitoring process variability.*
[15] Lim HA and Midi H 2016 Diagnostic robust generalized potential based on index set euality (DRGP (ISE)) for the identification of high leverage points in linear model *Comput. Stat.* **31** 859–77.
[16] Abd Mutalib SSS, Satari SZ and Wan Yusoff WNS 2019 A New Robust Estimator to Detect Outliers for Multivariate Data *J. Phys. Conf. Ser.* **1366** 1 - 9.
[17] Rencher AC 2002 *Methods of Multivariate Analysis* (John Wiley & Sons, Inc.).
[18] Pan J-X, Fung W-K and Fang K-T 2000 Multiple outlier detection in multivariate data using projection pursuit techniques *J. Stat. Plan. Inference*. **83** 153–67.
[19] Hadi AS 1992 Identifying multiple outliers in multivariate data *J. R. Stat. Soc. Series B Stat. Methodol.* **54** 761–71.
[20] Rocke DM and Woodruff DL 1996 Identification of outliers in multivariate data *J. Am. Stat. Assoc.* **91** 1047–61.
[21] Kosinski AS 1998 A procedure for the detection of multivariate outliers *Comput. Stat. Data Anal.* **29** 145–61.
[22] Becker C and Gather U 1999 The masking breakdown point of multivariate outlier identification rules *J. Am. Stat. Assoc.* **94** 947–55.
[23] Atkinson AC and Mulira HM 1993 The stalactite plot for the detection of multivariate outliers *Stat Comput.* **3** 27–35.
[24] Wang K and Lan H 2020 Robust support vector data description for novelty detection with contaminated data *Eng. Appl. Artif. Intell.* **91** 1 - 10.
[25] Affindi AN, Ahmad S and Mohamad M 2019 A comparative study between ridge MM and ridge least trimmed squares estimators in handling multicollinearity and outliers *J Phys Conf Ser.* **1366**.
[26] Midi H, Hendi HT, Arasan J and Uraibi H 2020 Fast and robust diagnostic technique for the

detection of high leverage points *Pertanika J. Sci. Technol.* **28** 1203–20.
[27] Sajana OK and Sajesh TA 2018 Detection of multidimensional outlier using multivariate spatial median *J. Comput. Math. Sci.* **9** 1875–81.
[28] Kunjunni SO and Abraham ST 2020 Multidimensional outlier detection and robust estimation using Sn covariance *Commun. Stat. Simul. Comput.* 1–11.
[29] Atkinson AC 1994 Fast very robust methods for the detection of multiple outliers *J. Am. Stat. Assoc.* **89** 1329–39.
[30] Jerison HJ 1973 *Evolution of the brain and intelligence* (Academic Press Inc.).
[31] Fung WK 1993 Unmasking outliers and leverage points: A confirmation *J. Am. Stat. Assoc.* **88** 515–9.
[32] Hawkins DM, Bradu D and Kass G V 1984 Location of several outliers in multiple-regression data using elemental sets *Technometrics* **26** 197–208.
[33] Maronna RA and Yohai VJ 1995 The behavior of the Stahel-Donoho robust multivariate estimator *J. Am. Stat. Assoc.* **90** 330–41.
[34] Daudin JJ, Duby CD and Trecourt P 1988 Stability of principal component analysis studied by the bootstrap method *Stat. A J. Theor. Appl. Stat.* **19** 241–58.